UNITED STATES PATENT APPLICATION

for

A Method for Manufacturing High Density Flash Memory and High Performance Logic on a Single Die

Inventors:

Henry S. Chao

Ervin T. Hill

Prepared by:

BLAKELY, SOKOLOFF, TAYLOR & ZAFMAN
12400 Wilshire Boulevard
Seventh Floor
Los Angeles, CA 90025-1026
(408) 720-8300

Attorney Docket No.: 042390.P17266

"Express Mail" mailing label number: EV 336 583 417US
Date of Deposit: November 10, 2003
I hereby certify that I am causing this paper or fee to be deposited with the United States Postal Service "Express Mail Post Office to Addressee" service on the date indicated above and that this paper or fee has been addressed to the Assistant Commissioner for Patents, Washington, D. C. 20231
Teresa Edwards
(~~Typed or~~ printed name of person mailing paper or fee)
Teresa Edwards
(Signature of person mailing paper or fee)
November 10, 2003
(Date signed)

A METHOD FOR MANUFACTURING HIGH DENSITY FLASH MEMORY AND HIGH PERFORMANCE LOGIC ON A SINGLE DIE

BACKGROUND OF THE INVENTION

1. FIELD OF THE INVENTION

[0001] The present invention relates to the field of semiconductor device fabrication, and in particular, to fabricating both high density flash memory transistors and high performance logic transistors on the same semiconductor die.

2. DISCUSSION OF RELATED ART

[0002] Present semiconductor fabrication process technology may be used to generate discrete high density flash memory and high performance logic on separate wafers, and hence, separate chips. These two types of devices are not easily manufactured on the same wafer or on the same die.

[0003] Figure 1 illustrates a logic transistor 100. Logic transistor 100 is formed on substrate 102. The gate stack of logic transistor 100 consists of a gate electrode layer 106 deposited over the gate dielectric 104 on substrate 102. Source/drain spacer liner dielectric 108 is formed on either side of the gate stack. Source/drain spacer dielectric 110 is formed on either side of the gate stack on top of the source/drain spacer liner dielectric 108.

[0004] Figure 2 illustrates a flash memory transistor 200. Flash memory transistor 200 is formed on substrate 102. The gate stack of flash memory transistor 200 consists of a control gate electrode layer 113 deposited over an inter-electrode dielectric 112, over a floating gate layer 111, over the gate dielectric 104, on a substrate 102. Source/drain spacer liner dielectric 108 is formed on either side of the flash memory gate stack. Source/drain spacer

dielectric 110 is formed on either side of the gate stack on top of the source/drain spacer liner dielectric 108.

[0005] Because the flash memory gate stack is significantly different from the logic gate stack, separate gate patterning masking steps are required to enable the different etches required to etch each gate stack. However, patterning of one set of gates will create large topographic steps at the boundaries between the flash and logic regions, because the height of the flash gate stack is greater than the height of the logic gate stack. The topographic steps may lead to large lithographic proximity effects where there is poor control over the critical dimensions near the flash/logic boundaries.

[0006] Presently, both flash memory and logic transistors may be formed on the same die, however present semiconductor fabrication methods do not permit fabrication of both flash memory transistors having minimum critical dimensions and logic transistors having minimum critical dimensions on the same die. Large topographic steps formed between the flash gate stack and the logic gate stack lead to poor control over gate critical dimensions, thus only one of either flash or logic gates having minimum critical dimensions may be formed on a die.

BRIEF DESCRIPTION OF THE DRAWINGS

[0007] Figure 1 is an illustration of a cross-sectional view of a logic transistor.

[0008] Figure 2 is an illustration of a cross-sectional view of a flash memory transistor.

[0009] Figure 3 is a flow diagram illustrating a process in accordance with one embodiment of the present invention.

[0010] Figure 4 is an illustration of a cross-sectional view of a logic gate stack and a flash memory gate stack on the same substrate.

[0011] Figure 5 is an illustration of a cross-sectional view of a logic gate stack and a flash memory gate stack on the same substrate after a hardmask layer is formed.

[0012] Figure 6 is an illustration of a cross-sectional view of a logic gate stack and a flash memory gate stack on the same substrate after the hardmask is patterned in the logic region.

[0013] Figure 7 is an illustration of a cross-sectional view of a logic gate stack and a flash memory gate stack on the same substrate after the flash memory gate stack is patterned to form flash memory gates.

[0014] Figure 8 is an illustration of a cross-sectional view of a logic gate stack and a flash memory gate stack on the same substrate after the logic gate stack has been etched.

[0015] Figure 9 is an illustration of a cross-sectional view of a logic gate stack and a flash memory gate stack on the same substrate after the flash memory gate stack has been partially etched.

[0016] Figure 10 is an illustration of a cross-sectional view of a logic gate stack and a flash memory gate stack on the same substrate after both gate stacks have been etched.

DETAILED DESCRIPTION OF THE PRESENT INVENTION

[0017] In the following description, numerous specific details are set forth, such as exact process steps, in order to provide a thorough understanding of the present invention. It will be apparent, however, to one skilled in the art that these specific details need not be employed to practice the present invention. In other instances, well known components or methods have not been described in detail in order to avoid unnecessarily obscuring the present invention.

[0018] Figure 3 is a flow diagram 300, showing a process in accordance with one embodiment of the present invention. Flow diagram 300 illustrates a general method of forming high density flash memory transistors and high performance logic transistors on the same substrate. In one embodiment of the present invention, high density flash memory transistors may have gate lengths of less than 150 nm, and may be located less than 400 nm from each other. High performance logic transistors may have gate lengths of less than approximately 200nm. In one embodiment of the present invention, the high density flash memory and the high performance logic transistors are formed on the same semiconductor wafer. In another embodiment of the present invention, the high density flash memory and the high performance logic transistors are formed on the same die. According to an embodiment of this invention, the smallest line widths and pitches at a technology node for both flash and logic transistors can be generated on the same semiconductor substrate.

[0019] First, a flash memory gate stack and a logic gate stack are formed on a substrate, as set forth in block 310. Conventional semiconductor processing techniques may be used to form each gate stack, as described below. After the gate stacks have been deposited, a hardmask layer is deposited over both the flash gate stack and the logic gate stack, as set forth in block 320. The hardmask layer is then patterned in the logic region, as set forth in block 330. Next, the flash memory gates are patterned, as set forth in block 340. Finally,

the logic gate stack is etched, using the remaining hardmask layer as a mask, as set forth in block 350.

[0020] Each block of the flow diagram is illustrated in Figures 4-10, and will be described in more detail below.

[0021] Figure 4 illustrates a logic gate stack 130 and a flash memory transistor gate stack 140 formed adjacent to each other or in close proximity to one another on a single substrate 102 according to one embodiment of the present invention. The logic gate stack 130 may be comprised of a gate electrode layer 106 deposited over the gate dielectric 104 on semiconductor substrate 102, and is formed in logic region 400. In one embodiment of the present invention, the gate dielectric 104 may comprise a silicon oxide or a silicon oxynitride. The gate electrode layer 106 may comprise polysilicon. The flash memory gate stack 140 may be comprised of a control gate electrode layer 113 formed over an inter-electrode dielectric 112, over a floating gate layer 111, over the gate dielectric 104, on semiconductor substrate 102, and is formed in flash region 410. In one embodiment of the present invention, the gate dielectric 104 may comprise a silicon oxide or silicon oxynitride. The inter-electrode dielectric 112 may comprise an ONO (oxide-nitride-oxide) layer. The floating gate layer 111 may comprise an n-type polysilicon layer. The control gate electrode 113 may comprise polysilicon. The control gate electrode 113 of the flash gate stack is the same layer as the gate electrode layer 106 of the logic gate stack, thus control gate electrode 113 and gate electrode 106 comprise the same material. In one embodiment of the present invention, the logic gate stack is approximately 2000Å thick. In one embodiment of the present invention, the flash gate stack is approximately 2500 Å thick. The thickness of each gate stack may vary considerably depending on the number and thickness of the layers comprising the gate stack.

[0022] In one embodiment of the present invention, the flash and logic gate stacks are formed in close proximity to one another on a single semiconductor substrate in the following manner. First the isolation is formed. The gate oxide is then grown or deposited in both the flash and the logic regions. After forming the gate oxide, the flash floating gate

layer is deposited on both the flash and the logic regions. The flash floating gate is then partially defined in the flash region and is removed from the logic region. The flash floating gate may be self aligned to the flash array isolation, or may be defined by a patterning step. After defining the flash floating gate, an interlayer dielectric is deposited in both the flash and the logic regions. The interlayer dielectric and any other dielectric, including the gate oxide previously formed, are removed from the logic region. Next, the logic gate dielectric is grown, and the control gate is deposited in both the flash and the logic regions. In other embodiments of the present invention, a different method may be used to form the flash and logic gate stacks in close proximity to one another on the same die.

[0023] The flash memory gate stack 140 is thicker than the logic gate stack 130, because the flash memory gate stack may include at least two additional layers not found in a logic gate: the inter-electrode dielectric and the floating gate. The height difference between the flash memory gate stack and the logic gate stack is typically between 500Å to 1000Å, but may be greater or less than this range depending on the composition of each of the respective gate stacks.

[0024] Figure 5 illustrates a hardmask layer 114 formed over both the logic gate stack 130 and the flash gate stack 140. In one embodiment of the present invention, the hardmask layer 114 is an anti-reflective coating (ARC) hardmask material. The ARC hardmask layer serves to eliminate undesirable optical effects from subsequent lithography operations. Ideally, the hardmask will comprise a material that can be selectively etched with respect to the gate stack layers. The ARC hardmask layer may be comprised of an oxide/oxynitride film stack. In other embodiments, the ARC hardmask layer may be comprised of nitride, carbon, or combinations of different anti-reflective coatings. The thickness of the hardmask layer and the properties of the layer are determined for optimal lithography performance in subsequent process operations. Ideally, the hardmask layer must be targeted to meet optimal lithography variability while being thick enough to withstand multiple etches and thin enough so that it may be reasonably easily removed. In one embodiment of the

present invention, for a 248nm lithography wavelength, the hardmask layer is between 200 and 400Å thick. In one embodiment of the present invention, the material used for the hardmask layer will withstand subsequent process operations without significant degradation.

[0025] After hardmask layer 114 is formed, the hardmask layer 114 in the logic region 400 of the die is patterned using standard lithographic techniques, and etching to form hardmask 115 as shown in Figure 6. First, a layer of resist 116 is formed on the top surface of the die. In the logic region, the hardmask is patterned to define the regions where logic gates are desired. In one embodiment of the present invention, the hardmask is removed by an etch process. The remaining regions of hardmask 115 define the areas where logic transistor gates will be formed. The flash region 410 remains covered with resist while the hardmask is patterned in the logic region. After patterning the hardmask 115 in logic region 400, resist 116 is removed from the surface of the die.

[0026] Next, the flash memory gates in the flash region of the die 410 are patterned using standard lithographic techniques, as illustrated in Figure 7. A layer of resist 118 is formed on the top surface of the die in both the flash and logic regions. The flash gates are then patterned in the flash region 410. The topographic step, S, between the logic and flash regions of the die is minimized because the logic gate electrode has not yet been completely etched. In one embodiment of the present invention, the topographic step, S is approximately 500Å. Thus, lithographic resist thickness variations are minimized at the logic/flash interface, which in turn reduces variations in critical dimensions near the logic/flash interface. The hardmask layer also contributes to further reduction of critical dimension variation due to any residual resist variations.

[0027] In one embodiment of the present invention, the entire flash gate stack is patterned, including the hardmask 114, gate electrodes 106, inter-electrode dielectric 112, and gate dielectric 104, as illustrated in Figure 7.

[0028] The logic region 400 of the die remains covered with resist while all or part of the flash gate stack is patterned. After patterning the flash gate stack as described above, resist 118 is removed from the surface of the die.

[0029] Finally, as illustrated in Figure 8, the remaining hardmask 115 in the logic region is used as a mask to etch the logic gate electrode 106, forming logic gates. In one embodiment of the present invention, a layer of resist is patterned to cover the flash memory gate stack with resist 120, leaving only the logic region 400 exposed. The logic gate stack is then etched using the remaining hardmask as a pattern, and regions of the logic gate stack that are not protected by hardmask 115 are removed. The resist 120 protects the flash memory gate stack while the logic gate stack is etched, preventing degradation of the gate stack and substrate in the flash region. This is a non-critical masking operation. After the logic gate stack is etched, resist 120 is removed from the surface of the die.

[0030] In another embodiment of the present invention, the both the flash region 410 and the logic region 400 may be exposed while the logic gate stack is etched, and no resist is used to protect the flash region 410. In this embodiment, a highly selective etch process must be used such that only the logic gate electrode material will be etched, to ensure that no substrate pitting occurs in the flash region 410. Because the hardmask layer in both the flash and the logic regions is exposed during etching, the hardmask layer may be sufficiently damaged by the final logic gate stack etch so that it may subsequently be more easily removed from the surface of the die.

[0031] Figure 9 illustrates another embodiment of the present invention. In this embodiment only the hardmask 114, control gate 106, and inter-electrode dielectric 112 of the flash gate stack are etched during the flash gate patterning operation as described above. The floating gate electrode layer 111 and the gate dielectric layer 104 are left intact. The remaining floating gate electrode and gate dielectric layer may be removed in a subsequent process operation to complete the formation of the flash gate. In one embodiment of the present invention, the remaining flash floating gate electrode is etched at the same time the logic gate electrode is etched.

[0032] In another embodiment of the present invention, both the flash region 410 and the logic region 400 are exposed, and both the logic gate stack and the remaining flash floating gate electrode are etched, using the hardmask in the logic region and the hardmask in the flash region as a mask for the etch.

[0033] After both the logic gate(s) 150 and the flash gate(s) 160 have been formed, the hardmask may be removed from to top surface of the logic gate(s) and the flash gate(s), as illustrated in Figure 10. Removal of the hardmask allows for complete silicidation of the gate electrode in subsequent processing operations.

[0034] In one embodiment of the present invention, the gate length of the high performance logic transistor gate, Lg_{logic}, is less than approximately 150nm. The gate length of the high density flash transistor gate, Lg_{flash}, is less than approximately 200nm. The high density flash transistor pitch is less than 400 nm.

[0035] Conventional processing techniques may be used to complete transistor formation for both the flash memory transistors and the logic transistors, including, but not limited to silicidation of the gate electrode, formation of source/drain regions, formation of source/drain spacer liner dielectrics, and formation of source/drain spacer dielectrics.

[0036] The present invention may be implemented with various changes and substitutions to the illustrated embodiments. For example, the present invention may be implemented on various types of thin film stacks having different heights. The present invention is not limited only to flash memory and logic transistor gate stacks. Furthermore, the present invention may be implemented on flash memory gates and logic gates whose gate stacks vary from those described herein. For example, a flash memory gate stack or a logic gate stack may contain additional or different layers than those described herein.

[0037] Although specific embodiments, including specific parameters, methods, and materials have been described, it will be readily understood by those skilled in the art and having the benefit of this disclosure, that various other changes in the details, materials, and arrangements of the materials and steps which have been described and illustrated in

order to explain the nature of this invention may be made without departing from the principles and scope of this invention as expressed in the subjoined claims.